Text Zoning and Classification for Job Advertisements in German, French and English

Gnehm, Ann-Sophie ; Clematide, Simon

Abstract: We present experiments to structure job ads into text zones and classify them into professions, industries and management functions, thereby facilitating social science analyses on labor marked demand. Our main contribution are empirical findings on the benefits of contextualized embeddings and the potential of multi-task models for this purpose. With contextualized in-domain embeddings in BiLSTM-CRF models, we reach an accuracy of 91% for token-level text zoning and outperform previous approaches. A multi-tasking BERT model performs well for our classification tasks. We further compare transfer approaches for our multilingual data.

DOI: https://doi.org/10.18653/v1/2020.nlpcss-1.10

Originally published at:
Gnehm, Ann-Sophie; Clematide, Simon (2020). Text Zoning and Classification for Job Advertisements in German, French and English. In: Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science, Online, 1 November 2020 - 30 November 2020. Association for Computational Linguistics, 83-93.
DOI: https://doi.org/10.18653/v1/2020.nlpcss-1.10

Text Zoning and Classification for Job Advertisements in German, French and English

Ann-Sophie Gnehm
Institute of Sociology
University of Zurich
gnehm@soziologie.uzh.ch

Simon Clematide
Department of Computational Linguistics
University of Zurich
simon.clematide@uzh.ch

Abstract

We present experiments to structure job ads into text zones and classify them into professions, industries and management functions, thereby facilitating social science analyses on labor marked demand. Our main contribution are empirical findings on the benefits of contextualized embeddings and the potential of multi-task models for this purpose. With contextualized in-domain embeddings in BiLSTM-CRF models, we reach an accuracy of 91% for token-level text zoning and outperform previous approaches. A multi-tasking BERT model performs well for our classification tasks. We further compare transfer approaches for our multilingual data.

1 Introduction

Text mining on job advertisements has become important to analyze labor market demand, since job ads provide unique job-level data on employers' staff needs (Atalay et al., 2020; Das et al., 2020; Calanca et al., 2019; Dawson et al., 2019). Our proposed techniques will be useful to study how job tasks and skill demand have developed in Switzerland in different labor market segments over the last decades. We present preparatory work for precise skill and task extraction: First, we structure job ads into text zones, that is, text parts dedicated to particular topics. Second, we classify job ads into professions, industries and management function. By replacing human annotation with scalable NLP, more fine-grained analyses on big data will be feasible.

Job ads contain information on topics such as the company, the job, or required qualifications. For an accurate extraction of skills and tasks, we need to identify the corresponding text zones, as many key terms are ambiguous, for instance 'dynamic' might refer to a personality trait or to a dynamic CRM system. Information on different topics can be densely packed in sentences, thus it seems most reasonable to formalize text zoning for job ads as token-level sequence labeling. In addition to this structuring of job ads, we need automatic classifications of job ads to enable detailed analysis, most importantly into professions, but also into industries and management functions.

For Swiss data in German, French, English and Italian we need multilingual approaches. Most (labeled) data however is in German. To avoid sparse data problems for minority languages, we thus experiment with transfer approaches.

The empirical experiments presented a) investigate the benefit of contextualized embeddings for text zoning, b) compare multilingual modeling with machine translation based approaches, and c) explore the potential of multi-task models for sequence labeling and text classification.

2 Related Work

Gnehm (2018) achieved an accuracy of 89.8% for the text zoning task at hand, namely token-level sequence labeling of job ads into eight zones with BiLSTMs, task-specific word embeddings and ensembling.[1] Hermes and Schandock (2017) segment on paragraph level, distinguish four classes, and reach accuracy of 97% with KNN in a multi-label classification. Grüger and Schneider (2019) extract HTML lists for IT job ads. Distinguishing between four list classes, they reach accuracy of 95% with LinearSVC. These two less fine-grained approaches are not directly comparable to ours.

Classification of professions is often provided by companies, and their methods and performance are not reported (Burke et al., 2020; Das et al., 2020; Calanca et al., 2019). Atalay et al. (2020) use embedding similarity measures to match jobs to 110 classes, and reach an accuracy of 53%.

[1] See Appendix A.3 for zone definitions and examples.

Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science, pages 83–93

https://doi.org/10.18653/v1/P17

3 Experimental Data

We use two job ad data sets, differing in size and data collection method. Both have each advantages for experiments here and for future analyses.

The **Swiss Job Market Monitor (SJMM)**[2] corpus consists of 80,000 job ads in German, French, English and Italian, from yearly samples representative for the Swiss job market, back to 1950. High-quality human annotations of profession, industry, and management function are available for all job ads. Text zones are annotated on German job ads until 2014. The SJMM provides us hence with labeled data for supervised machine learning experiments, and will allow analyses of how job tasks and skills developed over the last decades.

The **Online Ads (OA)** corpus contains 9 million ads in German, French, English and Italian from job portals and company websites in Switzerland crawled since 2012 by a private company. This big data set seems valuable for building in-domain embeddings for our experiments, and makes fine-grained analyses feasible for future research.

Text zoning in the SJMM is operationalized as introduced in Gnehm (2018). Eight zones are distinguished based on their content, and the text is segmented on token level.[3] Token level seems most appropriate, as information on different zones can be densely packed in single sentences. Not every ad contains information on every zone (e.g. not every job ad specifies personality traits of the ideal candidate) and zone distribution is strongly skewed: The job description (z6) comprises with more than 30% the largest share of tokens, whereas the least frequent zone, reason of the vacancy (z2), comprises 0.5% of tokens. Tokens show high zone ambiguity, with more than 90% of tokens showing up in more than one zone.

In text zoning experiments, we use the data split of Gnehm (2018), for comparability: Aiming for a model optimized for future application, dev and test set (*test set A*) are restricted to each 10% (n=650) of the most recent available data (2010-2014), the remaining 80% and all data further back to 1970 (n=22,700) serve as training data. In pure text classification experiments, we can use all multilingual SJMM data from the time span of interest (1990-2018, n=34,600), and take 80% for train, and 10% for dev and test set each (*test set B*).[4]

[2]Available under forsbase.unil.ch (Buchmann et al., 2019)

[3]Definitions and examples are shown in Appendix A.3.

[4]See Appendix A.1 for distribution over languages.

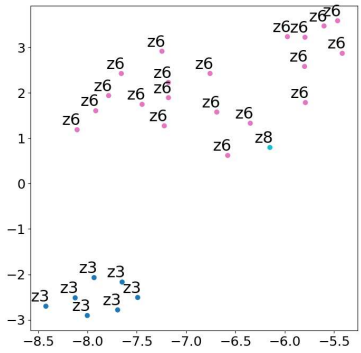

Figure 1: UMAP plot of FLAIRSJMM vectors for the term 'Ansprechpartner' *(contact person)* in test set A, appearing in zones job task (z6), wanted personality (z8) and residual (z3)

4 Experiments

4.1 Text Representation

We experiment with different text embeddings: **Static type-level** fastText (FT) embeddings provide a single vector for all occurrences of a word (Bojanowski et al., 2017). **Contextualized** embeddings allow to represent different word senses by capturing the semantics of surrounding text. We contrast BERT **sub-word** embeddings (Devlin et al., 2019) with **character-based** FLAIR embeddings (Akbik et al., 2018).

Given the large amount of in-domain texts, we train FLAIR embeddings on both of our corpora (FLAIRSJMM, FLAIROA).[5] We systematically compare the effect of these **in-domain** vs. **general domain** embeddings mentioned above.

Qualitative evidence for the usefulness of contextualized in-domain embeddings for text zoning is provided in Figure 1 with a UMAP (McInnes et al., 2018) vector visualization of the semantic space. The term 'Ansprechpartner' *(contact person)* in the zone for job description (z6) depicts that part of the job is to serve as contact person, probably for clients or co-workers, and the same term in the zone for the wanted personality (z8) hints furthermore that this person should be approachable or trustworthy. *Contact person* in the residual text (z3) however, simply refers to a contact information for the application procedure. The separation of the respective vectors in the semantic space in Figure 1 shows that such zone specific meanings can be recognized with our contextualized in-domain embeddings.

[5]Training hyper-parameters and preprocessing for all experiments can be found in the Appendix A.2.

Embeddings	Accuracy	Ens.
Gnehm (2018)$_d$	0.893 ± 0.001	0.898
FT$_g$	0.891	
FLAIR$_g$	0.902	
FLAIROA$_d$	0.907	
FLAIRSJMM$_d$	0.908 ± 0.001	0.909
FLAIRSJMM$_d$+FLAIROA$_d$	0.908	
FLAIRSJMM$_d$+FT$_g$	**0.909 ± 0.001**	**0.910**
BERT$_g$ first	0.880	
BERT$_g$ mean-pooling	0.881	
BERT$_g$ fine-tuned first	0.896	
BERT$_g$ fine-tuned mean-pooling	0.892	

Table 1: Accuracy of text zoning on test set A for different in-domain$_d$ and general$_g$ embeddings. The results with standard deviation report averages of 3 runs (of 5 runs for baseline by Gnehm (2018)) and column Ens. reports their majority vote ensemble (Rokach, 2010).

4.2 Text Zoning and Joint Classification

In this section, we first assess different text representations for our **text zoning** task, and evaluate the benefits of contextualized embeddings compared to previous work (Gnehm, 2018). We then explore the potential of **joint classification**, that is, including the classification of job ads into professions, industries and management functions in the sequence labeling text zoning task. Such a multi-tasking model would be most convenient in practical application. We assume furthermore that all these tasks are somewhat related and simultaneous learning could be beneficial.

For all these experiments we use the sequence labeling architecture proposed by Huang et al. (2015), a bidirectional LSTM with a CRF layer, implemented in the flair NLP library (Akbik et al., 2018). Model selection is based on dev set accuracy, and we evaluate on test set. For selected models we repeat the experiment three times and report mean performance and standard deviation (Reimers and Gurevych, 2017).

Text Zoning: For the first series of experiments, results in Table 1 show that models featuring contextualized FLAIR embeddings outperform all others in token-level sequence labeling of text zones. The best setting combines in-domain FLAIR embeddings with general domain FT word embeddings, reaching an ensemble accuracy of 0.91 and improving the baseline of Gnehm (2018) by more than 1 percentage point.[6]

This corresponds to earlier findings for PoS tagging and NER by Akbik et al. (2018). They hypothesize that type-level embeddings capture semantics that is complementary to the character-level features of FLAIR.

Interestingly, FLAIROA embeddings built from the much larger online corpus are less useful than FLAIRSJMM, which is probably due to the fact that the SJMM text zoning data consist for the most part of job ads in print media.

The lower performance of the pretrained German BERT might be explained by sub-tokenization issues. The many compound nouns and abbreviations of our special domain seem to cause problems for building meaningful entities to calculate embeddings over.[7] Using the mean of all sub-token embeddings for a token does not resolve this, but an improvement can be observed if we fine-tune the embeddings to the task.[8]

We tried ensemble combinations of models with different input embeddings (not shown), and of models with three runs (see Table 1) The best ensembles reach accuracy of 0.91, indicating limited variance between models. The lack of performance increase is convenient, as running a single classifier is easier than applying ensembles.

Joint classification: In this second series of experiments, we investigate if it is beneficial to combine the sequence labeling text zoning task with the classification of industry (11 classes), profession (34 classes) and management function (2 classes) in a single model. To answer this question, we add three special class tokens and their labels at the end of each job ad text. We focus on the best FLAIRSJMM+FT embeddings for text zoning, and assess adding model capacity (layers, hidden states). To direct the model towards learning predictions for the three special tokens, we experimentally increase their class weights $w \in \{10, 50, 100\}$ in the loss function. For technical reasons, this is only applicable on models without CRF layer, hence we also assess the effect of CRF.

Different joint models in Table 2 show relatively stable results for text zoning, industry and management function classification, whereas for the more fine-grained profession classification, accuracy depends on the model specifics. Dropping CRF affects accuracy for industry, profession and management function strongly. This shows the interdependence of the three variables represented as neighboring tokens.

[6] See Appendix A.4 for per-class results.

[7] Subtokenizaton produces fragmented results for compound nouns ('Anforderung', '##spro', '##fil', *required profile*) and diploma abbreviations ('N', '##DS').

[8] A scalar layer mix (Liu et al., 2019a) does not help.

Model	Zoning	Ind.	Prof.	Mgmt.
FLAIRSJMM+FT, 1 hidden layer, hidden size 256	**0.910**	0.813	0.653	**0.952**
-CRF	0.909	0.634	0.530	0.864
+ size (512), special weights (*50) (-CRF)	0.902	0.831	0.696	0.914
+ size (512), special weights (*100) (-CRF)	0.898	**0.834**	0.695	0.918
+ layer (2), size (512), special weights (*50) (-CRF)	0.904	0.832	**0.717**	0.922
+ layer (2), size (512), special weights (*100) (-CRF)	0.901	**0.834**	0.706	0.923

Table 2: Accuracy of joint prediction sequence taggers for zoning (8 classes), profession (34 classes), industry (11 classes) and management function (2 classes) on test set A. Only a subset of all tested combinations is shown.

Large class weights $w \in \{50, 100\}$ compensate for this performance drop and tune the model to the fine-grained classification tasks. More capacity in the form of larger hidden sizes and additional layers is useful, although the second layer helps only in combination with other factors.[9] By adding model capacity and weighted loss of 50 for the classification tasks, we find the model that performs best regarding profession classification, with relatively good results for all other tasks.

4.3 Text Classification

Results in Section 4.2 suggest that simultaneous learning of profession, industry and management function classification might be beneficial, but not enough model capacity is devoted to these tasks when including them into sequence labeling. Therefore, we experiment in the following with multi-tasking text classification for these three tasks. In **monolingual experiments**, we assess different multi-tasking models for classification of profession, industry and management function, and benchmark them with respective single task models. At last, we conduct **multilingual experiments** for profession classification. The SJMM data set is multilingual, but most labeled data (75%) is available for German. Hence we test different transfer approaches to avoid sparse data problems.

With the text classification implementation by Flair (Akbik et al., 2018), we obtain document level representations for job ads by feeding FLAIR or FT embeddings into an RNN. For BERT embeddings, we take the topmost layer of the transformer model and fine-tune embeddings during training. Document embeddings are extracted from the '[CLS]' token. In both cases, actual class labels are calculated by a linear layer on top.

Monolingual Experiments: We compare single vs. multi-tasking classification models using the best embeddings from previous experiments.[10] In multi-tasking, we simultaneously predict profession (34 classes), industry (11 classes) and management function (2 classes). We feed each job ad once for each task into the data, adding each time a special token that specifies the task to learn.

Embeddings	Prof.	Ind.	Mgmt.
FLAIRSJMM+FT *sT*	0.765		
FLAIRSJMM+FT *mT*	0.756	0.806	**.931**
BERT *sT*	**0.778**	**0.819**	0.920
	± 0.005	± 0.007	± 0.001
BERT *mT*	0.773	0.818	0.928
	± 0.004	± 0.003	± 0.003

Table 3: Accuracy for profession (34 classes), industry (11 classes) and management function (2 classes) in single (*sT*) and multi-task (*mT*) classification on test set B

With text classifiers and BERT embeddings, we reach an accuracy of 0.778 for professions (see Table 3. Although test sets A and B are not directly comparable, this surpasses sequence labeling results. For the other, somewhat less important variables, accuracy here is slightly lower.[11] BERT outperforms in multi- and single task classification our domain-specific contextualized embeddings, probably because BERT embeddings get fine-tuned to the task during training. Multi-tasking does not seriously alter profession classification, and the multi-tasking BERT reaches similar accuracy for industry and management function as single-task classifiers. It is thus reasonable to a go for the BERT multi-tasking classifier.

A detailed error analysis for professions further strengthens trust in the model. First, prediction probabilities and errors are strongly correlated: While for $p \geq 0.9$ error rate is only 12%, with $p \leq 0.5$ error rate is 75%. Thus, probabilities

[9] See ablation study in Table 11 in Appendix.

[10] FLAIRSJMM+FT and BERT performed best in classification of 11 professions. And, classification worked better on the whole job ad text than just on the job description (z6).

[11] A multi-tasking BERT model trained on data split A reaches an accuracy of 0.835, 0.731, and 0.921 for profession, industries and management function.

are useful for error detection. Second, a human post-evaluation of a random sample of 20 errors with $p \geq 0.9$ showed that only 10% of these errors are considered hard errors. In 90% of the cases, several class labels can be seen as correct options, and the model prediction is appropriate. This underlines that our model copes well with a sometimes ambiguous classification task.

Multilingual Experiments: On a classification task for 11 professions, we compare two approaches.[12] First, we use machine translation (MT) (DeepL) to translate French and English job ads to German, and apply a classifier trained for German. We test in this approach further, if familiarizing the classification model during training with partially awkward wording ('Translationese') helps, by including automatic translations in our train (and dev) set.[13] Second, we train multilingual classifiers on our German, French and English data with general-domain, multilingual FLAIR and BERT embeddings.

In the MT approach, accuracy decreases strongly, for French around 10, for English even up to 20 percentage points (see Table 4). One reason for the stronger effect in English is that class distribution differs from German.[14] Adding translated ads indeed helps, and raises accuracy by 9 points for English (BERT) and French (FLAIRSJMM+FT). Why French results vary more with FLAIRSJMM+FT and English results more with BERT needs further investigation.

Best performing are multilingual BERT for French (0.744), and BERT with Translationese for English (0.693). Multilingual models are a convenient solution, because no MT is needed for their application. For the MT approach, including translated ads in training seems necessary, especially if class distributions differ between languages. Either way, due to being fine-tuned to the task, BERT outperforms our domain-specific FLAIR embeddings.[15]

Approach	Test set originally in: DE	FR	EN
MT & Model for DE:			
FLAIRSMM+FT	0.798	0.672	0.625
incl. Translationese		0.718	0.639
BERT	**0.811**	0.715	0.603
incl. Translationese		0.724	**0.693**
Multilingual Model:			
BERT	0.803	**0.744**	0.679
FLAIR	0.654	0.542	0.499

Table 4: Accuracy for profession (11 classes) for MT vs. multilingual approach on test set B

5 Conclusion

Contextualized embeddings facilitate precise information extraction. Our best single text zoning models with stacked in-domain FLAIR and general domain word embeddings outperform the baseline of Gnehm (2018) and reduce the relative error rate by 12%. The combination of sequence labeling for text zoning and text classification for professions, industries and management function in a single multi-task model did not lead to entirely satisfying results. But, we found a multi-tasking BERT text classifier that performs well and provides a convenient solution for structuring our corpus into professions, industries, and management function. Error analysis for profession classification raised trust in this model. The model's classification probabilities provide valuable information for post-validation and subsequent analyses.

Multilingual experiments showed that our classifiers are affected by MT. Utilizing translated material in training, or alternatively multilingual models, are potential strategies, but the question of the best transfer approaches for our multilingual data needs further investigation.

The most promising approach for future work seems to be the training of our own domain-specific BERT embeddings, both for optimizing classification and for intended subsequent skill and task extraction. This way, we can also exploit the large amount of data in the OA corpus. Another direction worthy to explore is multi-tasking, be it by including more variables, or by experimenting with more sophisticated approaches (Clark et al., 2019; Liu et al., 2019b).

Acknowledgments

We thank Dong Nguyen and the anonymous reviewers for their careful reading of this article and their helpful comments and suggestions, and Helen Buchs for her efforts in post-evaluation. This work is supported by the Swiss National Science Foundation under grant number 407740_187333.

[12] For the sake of sound evaluation, we choose here a broader classification scheme, and restrict experiments to French and English (the amount of ads in Italian is too small).

[13] Adding 4,100 (500) ads from French, 2,900 (350) from English to the original 20,700 (2,600) from German.

[14] See Table 13 in Appendix.

[15] The multilingual BERT without fine-tuning reaches accuracies below 0.3 for the 3 languages.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual String Embeddings for Sequence Labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA.

Enghin Atalay, Phai Phongthiengtham, Sebastian Sotelo, and Daniel Tannenbaum. 2020. The Evolution of Work in the United States. *American Economic Journal: Applied Economics*, 12(2):1–34.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Marlis Buchmann, Helen Buchs, Felix Busch, Ann-Sophie Gnehm, Urs Klarer, Jan Müller, Marianne Müller, Stefan Sacchi, Alexander Salvisbert, and Anna von Ow. 2019. *Stellenmarkt-Monitor Schweiz 1950 – 2018.* Soziologisches Institut der Universität Zürich.

Mary Burke, Alicia Sasser Modestino, Shahriar Sadighi, Rachel Sederberg, and Bledi Taska. 2020. No Longer Qualified? Changes in the Supply and Demand for Skills within Occupations. Federal Reserve Bank of Boston Research Department Working Papers, Federal Reserve Bank of Boston. Series: Federal Reserve Bank of Boston Research Department Working Papers.

Federica Calanca, Luiza Sayfullina, Lara Minkus, Claudia Wagner, and Eric Malmi. 2019. Responsible team players wanted: an analysis of soft skill requirements in job advertisements. *EPJ Data Science*, 8(1):1–20. Number: 1 Publisher: SpringerOpen.

Kevin Clark, Minh-Thang Luong, Urvashi Khandelwal, Christopher D. Manning, and Quoc V. Le. 2019. BAM! Born-Again Multi-Task Networks for Natural Language Understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5931–5937, Florence, Italy.

Subhro Das, Sebastian Steffen, Prabhat Reddy, Erik Brynjolfsson, and Martin Fleming. 2020. Forecasting Task-Shares and Characterizing Occupational Change across Industry Sectors. In *Harvard CRCS Workshop on AI for Social Good*.

N. Dawson, Marian-Andrei Rizoiu, Benjamin Johnston, and Mary-Anne Williams. 2019. Adaptively selecting occupations to detect skill shortages from online job ads. *2019 IEEE International Conference on Big Data (Big Data)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ann-Sophie Gnehm. 2018. Text Zoning for Job Advertisements with Bidirectional LSTMs. *Proceedings of the 3rd Swiss Text Analytics Conference - SwissText 2018, CEUR Workshop Proceedings*, 2226:66–74.

Joscha Grüger and Georg Schneider. 2019. Automated Analysis of Job Requirements for Computer Scientists in Online Job Advertisements:. In *Proceedings of the 15th International Conference on Web Information Systems and Technologies*, pages 226–233, Vienna, Austria. SCITEPRESS - Science and Technology Publications.

Juergen Hermes and Manuel Schandock. 2017. *Stellenanzeigenanalyse in der Qualifikationsentwicklungsforschung: Die Nutzung maschineller Lernverfahren zur Klassifikation von Textabschnitten.* Bundesinstitut fuer Berufsbildung, Bonn.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic Knowledge and Transferability of Contextual Representations. In *Proceedings of the 2019 Conference of the North*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019b. Multi-Task Deep Neural Networks for Natural Language Understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Leland McInnes, John Healy, and James Melville. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802.03426 [cs, stat]*. ArXiv: 1802.03426.

Nils Reimers and Iryna Gurevych. 2017. Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348, Copenhagen, Denmark. Association for Computational Linguistics.

Lior Rokach. 2010. Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1):1–39.

A Appendices

A.1 Data splits

	German
train set (1970-2014)	22,698
dev set (2010-2014)	672
test set A (2010-2014)	626

Table 5: Number of German SJMM job ads in data split A. Dev and test set restricted to years 2010-14

	German	French	English
train set	20,717	4,126	2,846
dev set	2,581	518	355
test set B	2,598	515	355

Table 6: Number of job ads in data split B for German, French and English, SJMM job ads from 1990-2018

A.2 Preprocessing & Training Parameters

Training parameters are set according to recommendations in the Flair library (Akbik et al., 2018) unless reported differently here.

Text representations: For FLAIRSJMM and FLAIROA we train forward and backward language models with LSTMs with one layer and 2048 hidden states on the SJMM (67MB) and the OA (4GB) corpus.

Preprocessing is kept simple: We map digits to 0, white space to single blanks, and replace web and e-mail addresses with special tokens (replaced-dns, replaced-email, replaced-url). We build our own domain-specific character dictionary, setting the rarest 0.0001% of characters to unknown.

We optimize with SGD, clip gradients at 0.25 and set dropout probability to 0.25. Sequence length is set to 250 and batch size to 100. We train our language models with a learning rate of 20 for 2 weeks, reaching perplexity of 1.73 (forward model), and 1.74 (backward model) for FLAIRSJMM and 1.45 and 1.46 on validation sets for forward and backward models of FLAIROA.

General-domain FLAIR embeddings are provided by Akbik et al. (2018), for German we use embeddings that are pretrained on a mixed corpus (Web, Wikipedia, Subtitles) and in the multilingual setting embedddings that are pretrained on JW300 corpus. For German BERT embeddings, we use the model trained by Deepset.ai with 12 layers, 768 hidden states, 12 heads and 110M parameters, for multilingual BERT embeddings a model with the same configurations, trained on cased text in 104 languages.

FT are German FastText embeddings without character feature provided in the Flair library.

Sequence labeling: We optimize with SGD, clipping gradients at 5. Minibatch size is 32 and training starts with learning rate of 0.1 and is annealed with factor 0.5 after 5 periods with no loss decrease. We stop training after 150 epochs, or as soon as the learning rate ≤ 0.0001. We use variational dropout ($p = 0.5$) and word dropout ($p = 0.05$) for regularization.

Text classification: For all models with FLAIR embeddings (FLAIRSJMM, FLAIRSJMM+FT, multilingual FLAIR), training parameters are as described above for sequence labeling. Classifier with German or multilingual BERT embeddings are optimized with Adam over 5 epochs, with a learning rate of 3.00E-05, in minibatches of 16.

A.3 Text Zoning: Definitions and Examples

For/z1 our/z1 attractive/z1
product/z1 portfolio/z1
we/z3 are/z3 looking/z3 for/z3
an/z6 interior/z6 designer/z6 ./z6
You/z3 offer/z3 :/z3
-/z7 solid/z7 vocational/z7 training/z7
and/z7 experience/z7 ,/z7
-/z8 creativity/z8 and/z8 versatility/z8 ,/z8
-/z8 ideally/z8 you/z8 are/z8
between/z8 25/z8 and/z8 40/z8
years/z8 old/z8 ./z8
We/z3 offer/z3 :/z3
-/z6 a/z6 high/z6 degree/z6
of/z6 autonomy/z6 ,z6
-/z6 a/z6 large/z6 studio/z6 ,/z6
-/z6 an/z6 interesting/z6 and/z6
stimulating/z6 permanent/z6 position/z6 ./z6
Please/z3 send/z3 your/z3 application/z3
to/z3 POC/z3 ,/z3 ADDR/z3 ,/z3
Foto/z1 Hobby/z1 Inc./z1

Table 7: Example of job ad with text zoning annotation (Gnehm, 2018), translated from German to English

zone	definition	example
z1	company description	'ein erfolgreiches Unternehmen der Baubranche' *'a successful company in the construction industry'*
z2	reason of vacancy	'für unsere neu eröffnete Filiale' *'for our newly opened branch'*
z3	administration & residual text	'Ihre Bewerbung senden Sie an' *'Please send your application to'*
z4	job agency description	'Ihr Partner für die Vermittlung von Dauerstellen' *'your competent partner for permanent position placements'*
z5	material incentives	'ansprechendes Salär' *'attractive salary'*
z6	job description	'für den Kundenempfang' *'for the customer reception'*
z7	required hard skills	'eine Ausbildung und Berufserfahrung als Sozialarbeiter' *'a degree and experience in social work'*
z8	required personality (soft skills)	'Sie sind diskret und belastbar' *'you are diplomatic and able to work under pressure'*

Table 8: Definitions and example of text zones (Gnehm, 2018), translations from German to English added to shortened examples

A.4 Detailed Results

	Frequency				
zone	abs.	rel.	precision	recall	F1
z1	22672	17.2%	0.898	0.921	0.909
z2	639	0.5%	0.863	0.757	0.807
z3	33186	25.2%	0.941	0.908	0.924
z4	964	0.7%	0.806	0.667	0.730
z5	2199	1.7%	0.870	0.752	0.807
z6	42610	32.4%	0.907	0.925	0.916
z7	16767	12.7%	0.917	0.925	0.921
z8	12515	9.5%	0.865	0.872	0.868

Table 9: Per zone frequencies, precision, recall and F1-values for best text zoning model FLAIRSJMM+FT on test set A, reaching accuracy of 0.91

prediction truth	z1	z2	z3	z4	z5	z6	z7	z8
z1	92.1% 20878	0.1% 29	1.8% 406	0.2% 56	0.3% 59	5.2% 1170	0.1% 16	0.3% 58
z2	6.4% 41	75.7% 484	7.4% 47	0.0% 0	0.0% 0	10.5% 67	0.0% 0	0.0% 0
z3	2.0% 654	0.0% 6	90.8% 30147	0.3% 97	0.1% 38	3.5% 1161	1.9% 638	1.3% 445
z4	23.2% 224	0.0% 0	8.4% 81	66.7% 643	0.0% 0	1.6% 15	0.0% 0	0.1% 1
z5	3.7% 81	0.0% 0	3.4% 75	0.0% 0	75.2% 1653	17.2% 378	0.1% 2	0.5% 10
z6	3.1% 1319	0.1% 38	1.3% 559	0.0% 2	0.3% 113	92.5% 39433	0.9% 388	1.8% 758
z7	0.1% 15	0.0% 0	2.4% 399	0.0% 0	0.1% 15	2.3% 392	92.5% 15509	2.6% 437
z8	0.3% 39	0.0% 4	2.6% 331	0.0% 0	0.2% 22	6.8% 853	2.9% 359	87.2% 10907

Table 10: Confusion matrix for best zoning sequence model FLAIRSJMM+FT on test set A, reaching accuracy of 0.91, cells show row percentages and frequencies

Model	Zoning	Ind.	Prof.	Mgmt.
FLAIRSJMM, 1 hidden layer, hidden size 256	0.909	0.786	0.601	0.915
- CRF	0.909	0.634	0.530	0.864
+ size (512)	0.909	0.800	0.653	0.915
+ layer (2)	0.907	0.736	0.602	**0.926**
+ special weights (*10) (-CRF)	0.907	0.802	0.673	**0.926**
+ special weights (*50) (-CRF)	0.900	**0.808**	**0.695**	0.925
+ FT	**0.910**	0.813	0.653	0.925

Table 11: Accuracy of joint prediction sequence taggers for zoning (8 classes), profession (34 classes), industry (11 classes) and Mgmt. position (2 classes) on test set A

	Frequency		FLAIRSJMM+FT		BERT	
			mT	*sT*	*mT*	*sT*
Industry	abs.	rel.	F1	F1	F1	F1
unidentifiable	89	5.9%	0.515		0.545	0.558
Agriculture, private households	54	3.6%	0.855		0.851	0.816
Chemical, Food, Textile Industry	212	14.2%	0.713		0.739	0.733
MEM industries	260	17.4%	0.753		0.762	0.761
Construction	170	11.4%	0.763		0.785	0.777
Trade, Trasnsportation	503	33.6%	0.791		0.816	0.818
Hospitality, entertainment, pers. services	211	14.1%	0.786		0.830	0.834
Finance, Insurance	178	11.9%	0.898		0.896	0.916
Company services	228	15.2%	0.716		0.703	0.722
Public administration	240	16.0%	0.882		0.880	0.880
Education, Science, Health	453	30.3%	0.942		0.939	0.941
Profession						
Agricultural, forestry, fishery workers	27	0.9%	0.926	0.926	0.933	0.930
Food & luxury goods production workers	14	0.5%	0.923	0.929	0.922	0.858
Metal &machinery workers	85	2.8%	0.759	0.776	0.736	0.764
Electronics, watch making, automotive workers	73	2.4%	0.761	0.757	0.789	0.785
Wood, paper production workers	29	0.9%	0.772	0.759	0.803	0.782
Chemical, plastic production workers	13	0.4%	0.560	0.462	0.522	0.612
Textile production, printing, storage workers	39	1.3%	0.713	0.769	0.640	0.656
Engineers	111	3.6%	0.745	0.690	0.733	0.716
Technicians	72	2.3%	0.504	0.472	0.579	0.572
Technical drafting workers	18	0.6%	0.773	0.944	0.753	0.755
Technical workers	46	1.5%	0.315	0.333	0.286	0.338
Machine operators	9	0.3%	0.737	0.778	0.700	0.804
IT professionals	96	3.1%	0.804	0.812	0.830	0.818
Construction workers	130	4.2%	0.827	0.863	0.833	0.846
Commerce, Sales professions	373	12.2%	0.814	0.851	0.827	0.829
Marketing and tourism professionals	47	1.5%	0.450	0.383	0.564	0.530
Fiduciaries	43	1.4%	0.521	0.558	0.615	0.569
Transportation professions	54	1.8%	0.804	0.778	0.777	0.815
Post, Telecommunication workers	21	0.7%	0.581	0.476	0.628	0.677
Hospitality, housekeeping workers	145	4.7%	0.884	0.897	0.912	0.918
Cleaning, hygiene and personal care workers	90	2.9%	0.789	0.835	0.835	0.855
Entrepreneurs, directors, senior officials	190	6.2%	0.606	0.579	0.620	0.627
Merchants, administrative professions	316	10.3%	0.773	0.792	0.792	0.785
Banking, Insurance professions	88	2.9%	0.700	0.782	0.768	0.762
Security workers	17	0.6%	0.774	0.824	0.911	0.911
Legal professions	25	0.8%	0.816	0.833	0.852	0.801
Media professionals	21	0.7%	0.700	0.667	0.713	0.756
Artists	4	0.1%	0.667	0.750	0.736	0.814
Welfare, care, counseling professions	60	2.0%	0.810	0.850	0.795	0.802
Educational professions	100	3.3%	0.822	0.810	0.863	0.848
Humanities, social and natural science	20	0.7%	0.514	0.400	0.518	0.508
Medical, pharmaceutical professions	57	1.9%	0.891	0.945	0.906	0.931
Therapy and nursing professions	145	4.7%	0.907	0.924	0.902	0.923
unclassifiable workers	20	0.7%	0.176	0.250	0.242	0.319
management function						
no	2121	81.6%	0.958		0.956	0.952
yes	477	18.4%	0.799		0.809	0.759

Table 12: Class frequencies and F1-values for multi-(*mT*) and single task (*sT*) classifiers on test set B, for SJMM+FT and BERT. For BERT we report mean values over 3 training runs.

	DE			FR			EN		
Professional Class	abs. Freq.	rel. Freq.	F1	abs. Freq.	rel. Freq.	F1	abs. Freq.	rel. Freq.	F1
Industry & Transport	360	13.6%	0.809	75	14.6%	0.784	8	2.3%	0.182
Construction	205	7.7%	0.848	13	2.5%	0.696	0	0.0%	
Technology & Science	245	9.3%	0.748	43	8.3%	0.674	51	14.4%	0.730
IT	143	5.4%	0.872	25	4.9%	0.750	63	17.7%	0.806
Trade & Sales	195	7.4%	0.836	73	14.2%	0.787	19	5.4%	0.595
Office & Administration	120	4.5%	0.747	43	8.3%	0.605	29	8.2%	0.613
Financial & Fiduciary Services	409	15.5%	0.817	52	10.1%	0.796	64	18.0%	0.773
Management & Organisation	248	9.4%	0.591	42	8.2%	0.444	86	24.2%	0.625
Hospitality & Personal Services	209	7.9%	0.875	44	8.5%	0.830	4	1.1%	0.750
Health	243	9.2%	0.934	45	8.7%	0.932	6	1.7%	0.727
Teaching & Public Services	269	10.2%	0.868	60	11.7%	0.748	25	7.0%	0.604

Table 13: Class frequencies and F1-values for profession classification (11 classes) of best performing approach for German (German BERT), French (multilingual BERT), and English (Machine Translation & German BERT with Translationese), on test set B